Environ Health Perspect
DOI: 10.1289/EHP172

Supplemental Material

A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations

Lydiane Agier, Lützen Portengen, Marc Chadeau-Hyam, Xavier Basagaña, Lise Giorgis-Allemand, Valérie Siroux, Oliver Robinson, Jelle Vlaanderen, Juan R. González, Mark J. Nieuwenhuijsen, Paolo Vineis, Martine Vrijheid, Rémy Slama, and Roel Vermeulen

Table of Contents

the hypothetical situation in which we had a perfect model (oracle) to illustrate the limitations of having highly correlated data. Results are given for scenarios with k=0,1,2,3,5,10,25 true predictors.

Figure S2. Additional parameters characterizing the performances of the statistical methods for scenarios set 1. Model performances are summarized by n_B the number of variables selected by the method (A), the ratio n_B/k with k the number of true predictors (=n_B when $k = 0$) (B), the error variance σ^2 (C), the mean bias (all coefficient values being equal to 1) (D), the mean absolute bias computed over the true predictors (E) and over the other variables that are not true predictors (F). For each scenario defined by a number of true predictors varying from 0 to 25, statistics over the 100 runs are summarized by their mean (dot) and the variability of each statistics is summarized using the one standard error both ways from the average value (vertical dotted line). DSA: Deletion/substitution/addition; ENET: Elastic net; EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression; GUESS: Graphical Unit Evolutionary Stochastic Search; sPLS: Sparse partial least-squares; TP: True Predictors.

Figure S3. Performances of the statistical methods for scenarios set 1 using alternative multiple hypothesis testing corrections for the EWAS and EWAS-MLR methods: the Bonferroni (Bon), permutation (perm) and Benjamini and Hochberg (BH) corrections. For each scenario defined by a number of true predictors varying from 0 to 25, statistics over the 100 runs are summarized by their mean (dot) and the variability of each statistics is summarized using the one standard error both ways from the average value (vertical dotted line). EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression.

Figure S4. Performances of the statistical methods according to the amount of correlation between the true predictors. The full line connects the results for scenarios with correlations between the true predictors (in absolute values) in the [0,1] range (scenarios set 1); the dashed line in the [0,0.2] range (scenarios set 2) and the dotted line in the [0.5,1] range (scenarios set 3). For each scenario defined by a number of true predictors varying from 0 to 25 (to 10 for set 3, Σ not containing 25 exposures that are correlated at a level >0.5), statistics over the 100 runs are summarized by their mean (dot). DSA: Deletion/substitution/addition; ENET: Elastic net; EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression; GUESS: Graphical Unit Evolutionary Stochastic Search; sPLS: Sparse partial least-squares.

Figure S5. Performances of the statistical methods when deviating from the assumption of normally distributed exposures. The full line connects the results for scenarios set 1 (exposures generated from a normal distribution); the dotted line corresponds to the same scenarios being tested on exposure data bootstrapped from the INMA environmental data (scenarios set 6). For each scenario defined by a number of true predictors varying from 0 to 25, statistics over the 100 runs are summarized by their mean (dot). DSA: Deletion/substitution/addition; ENET: Elastic net; EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression; GUESS: Graphical Unit Evolutionary Stochastic Search; sPLS: Sparse partial least-squares.

Figure S6. Performances of the statistical methods for varying effect sizes among true predictors. The full line connects the results for scenarios set 1 (all effect sizes equal to 1); the dotted line corresponds to the same scenarios except with effect sizes generated from a uniform distribution in [0.5,1.5] (scenarios set 7). For each scenario defined by a number of true predictors varying from 0 to 25, statistics over the 100 runs are summarized by their mean (dot). DSA: Deletion/substitution/addition; ENET: Elastic net; EWAS: Environment-wide association study; EWAS-MLR: EWAS-Multiple Linear Regression; GUESS: Graphical Unit Evolutionary Stochastic Search; sPLS: Sparse partial least-squares.

Additional Files

Supplemental Code and Data Zip File

Supplemental Code and Data Zip File Index

Excel File Table S1: Matrix of all Pearson pairwise correlation coefficients in Σ, the semi-definite matrix the closest to the correlation matrix derived from the INMA cohort